STATISTICS FOR THE DATA SCIENCE

Part - 3

- POISSON DISTRIBUTION
- NORMAL / GAUSSIAN DISTRIBUTION
- UNIFORM DISTRIBUTION
- Z - SCORE
- CENTRAL LIMIT THEORAM
- ESTIMATOR
- HYPOTHESIS AND TESTING MECHANISIM
- P -VALE
- Z - TEST

#Value_freeContent

@Krishan kumar

* Pmf :-

$$P(k,n,p) = {}^{n}C_{k}\, p^{k} (1-p)^{n-k}$$

for $k = 0, 1, 2, 3, \ldots, n$ where

$$^{n}C_{k} = \frac{\lfloor n}{\lfloor k \, \lfloor n-k}$$

⇒ **Mean :-**

$$\text{Mean} = np$$

⇒ **Variance :-**

$$\text{Var} = npq$$

⇒ **Std**

$$\text{Std} = \sqrt{npq}$$

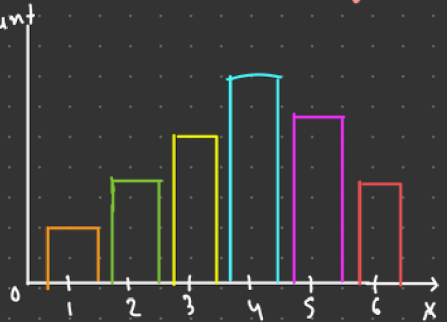
Poisson Distribution

* Discrete random variable. (Pmf)
* Discribe the number of event occuring in fixed time interval.

ex) ① No. of people visiting hospital every hour.

② No. of people visiting bank every hour.

Pmf



$\lambda = 3$, Expected no. of event occure at every time interval.

Means → How many people come at every hour.

Q What is the probability of a person to come at 5th hour.

Ans

$$Pmf = P(X=5) = \frac{e^{\lambda} - \lambda^{x}}{\lfloor x}, \quad \text{if } \lambda = 3 \text{ (औसत)}$$

$$= \frac{e^{3} - 3^{5}}{\lfloor 5} \Rightarrow 0.101 \rightarrow 10.1\%$$

there is 10% posibilities that at 5th hour 3 person come.

Q What is the probability to visit at 5th hour or 4th hour.

Ans

Pmf $P(X=5) + P(X=4)$

* Mean of poission distribution

Mean $= \Sigma(x) = \mu$

$\mu = \lambda * t$

λ = expected no. of event occur at every time interval

t = time interval

Variance of poission = $\lambda * t$

Normal / Gaussian Distribution (PDF)

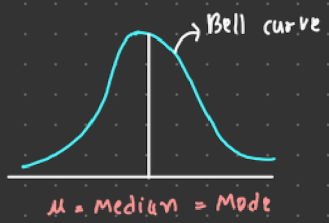
Notation :- $N(\mu, \sigma^2)$

Parameter :-

$\mu \in R$ (Mean) — Real number

$\sigma^2 \in R > 0$ = Variance

$x \in R$ = Data points



$$PDF = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma}}$$

⇒ **Mean of normal distribution**

Mean = μ = Average

⇒ **Variance**

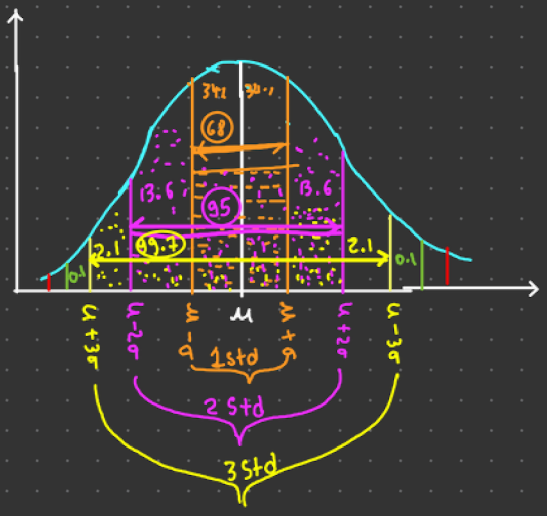
Var = σ^2

⇒ **Std**

$\sigma = \sqrt{Var}$

This role follow

68 - 95 - 99.7 % Rule



$P(\mu-\sigma \leq x \leq \mu+\sigma) \approx 68\%$

$P(\mu-2\sigma \leq x \leq \mu+2\sigma) \approx 95\%$

$P(\mu-3\sigma \leq x \leq \mu+3\sigma) \approx 99.7\%$

Ex → ① Weight of the students in the class.
② Height of " " " " ".
③ IRIS Dataset [Sepal widht]

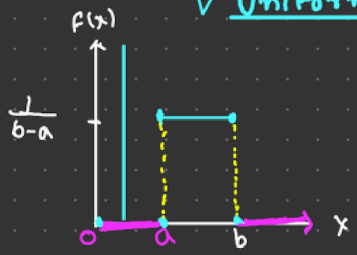
Uniform Distribution

(A) Continous uniform distribution (PdF)
(B) Discrete Uniform distribution (Pmf)

(A) Continous Uniform Distribution

In statistic, the continous uniform dis. OR Rectangular dis. is a family of symmetric probability dis. The dis discribe an experiment where there is an arbitrary outcome that lie between certain bounds.
the bounds are defined by the parametter a and b, which are the minimum and maximum va

V Uniform (PdF)

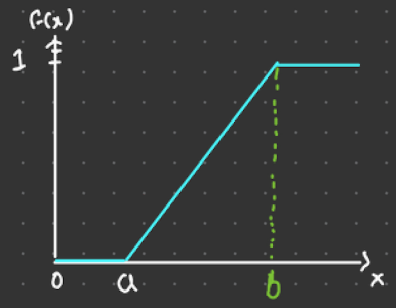


Notation → V(a,b)

Parametter →

$-\infty < a < b < \infty$

- **CDF**



$$pdf \rightarrow \begin{cases} \frac{1}{b-a} & \text{for} \quad x \in [a,b] \\ 0 & \text{, otherwise} \end{cases}$$

$$cdf \rightarrow \begin{cases} 0 & \text{for} \quad x < a \\ \frac{x-a}{b-a} & \text{for} \quad x \in [a,b] \\ 1 & \text{for} \quad x > b \end{cases}$$

Mean $= \frac{1}{2}(a+b)$

Variance $\rightarrow \frac{1}{12}(b-a)^2$

Median $\rightarrow \frac{1}{2}(a+b)$

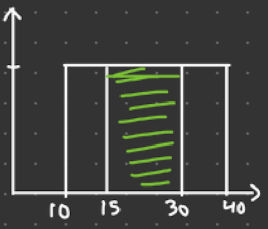
Ex → The no. of candie sold daily at a shop is uniformally dis. with a maxi at 40 and mini. of 10.

① Probability of daily Sales to falls between 15 and 30.

Sol →

$\frac{1}{b-a} \Rightarrow \frac{1}{40-10}$

$\Rightarrow \frac{1}{30}$



$x_1 = 15$

$x_2 = 30$

$$P(15 \leq x \leq 30) = (x_2 - x_1) \times \frac{1}{b-a}$$

$\Rightarrow 15 \times \frac{1}{30} \Rightarrow \frac{1}{2} \Rightarrow 0.5$

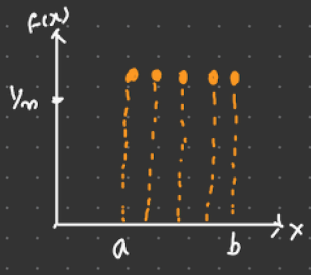
② $P(x \geq 20)$

Sol. $(40-20) \times \frac{1}{30}$

$\Rightarrow 0.66 \rightarrow$ 66%

(B) Discrete uniform dis. (Pmf)

In statistic the discrete uniform dis. is symmetric probability dis. wherin a finite number of value are equally likely to be observed every on of n value has equally probability 1/n. Another way of saying that "discrete uniform Dis" would be a known, finite number of outcome equally likely to happen.



Ex Rolling a dice

[1,2,3,4,5,6]

↳ P(1) → 1/6

⋮

P(6) → 1/6

a → 1

b → 6

Natation → $u(a,b)$

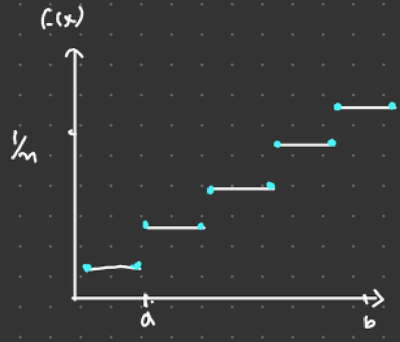
$\frac{1}{n} = n = b - a + 1$

Parametter → a, b with $b \geq a$

Pmf $= \frac{1}{n}$

Mean $= \frac{a+b}{2}$

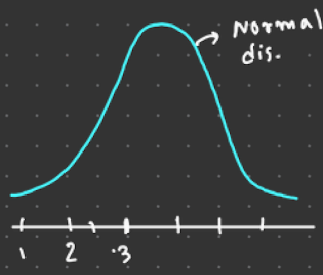
Medion ↗



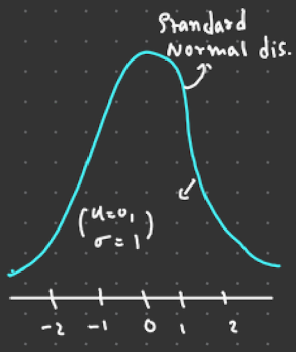
Standard Normal Distribution and Z-score

$X \in [1,2,3,4,5]$, $\mu = 3$

$\sigma = 1.414 \simeq 1$



Transition Technique → When $\mu = 0$, $\sigma = 1$



⇒ Z-Score

Z-score is used for the help for Normal dis. to Transform in standard normal distribution.

$$\text{Z-Score} = \frac{x_i - \mu}{\sigma}$$

$x_i = 1$, $\frac{1-3}{1} \Rightarrow -2$

$x_i = 2$, $\frac{2-3}{1} \Rightarrow -1$

$x_i = 3$, $\frac{3-3}{1} \Rightarrow 0$

$x_i \to 4 = \frac{4-3}{1} \to 1$

$x_i \to 5 \Rightarrow \frac{5-3}{1} \to 2$

Q Away from the mean for a specific mean how much Std fall for the 4.

↳ by Z-score → $x_i = 4 \Rightarrow \frac{4-3}{1} \to \boxed{1}$

Note :→ one std to the right → when value is +ve

one std to the left → when value is -ve

$\{\mu = 4, \sigma = 1\}$ (Normal dis. fun.)

Q How many standard deviation 4.5 is away from the mean?

Ans → $x_1 = 4.5$

Z-Score → $\frac{4.5 - 4}{1}$ → 0.5

0.5 Std from the right.

Q What percentage of data is falling above 4.5

$\mu \to 4$

$\sigma \to 1$

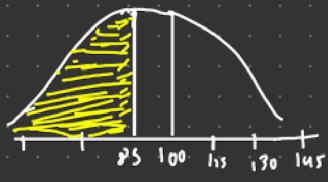
Z-Score → $\frac{2.5 - 4}{1} \Rightarrow 2.5$

Area under the curve (≤ 2.5) ⇒ 1 − 0.6681 → 6.6%

Problem → In india average IQ is 100, with a std of 15 what is the percentage of the population would you expact to have an IQ lower then 85.

$\mu \to 100$, $\sigma = 15$

Z-Score → $\frac{85 - 100}{15} \to \frac{-15}{15} \to -1$



Area under the curve = (≤ 85) = 0.15866 ⇒ 15.8%

Central limit theorem,

type, ① $X \simeq N(\mu, \sigma)$

$n = 20$

Sample distribution

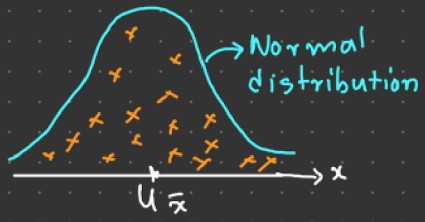
$S_1 \rightarrow [x_1, x_2, \ldots, x_{20}] = \bar{x}_1$

$S_2 \rightarrow [x_2, x_3, \ldots, x_{20}] = \bar{x}_2$

$\vdots$

$S_n \rightarrow \bar{x}_n \Rightarrow$ Sample mean.

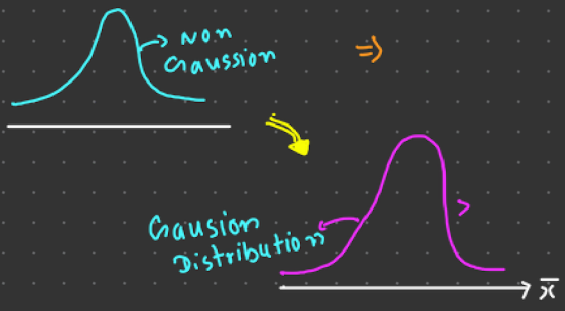
$\bar{x} = \{\bar{x}_1 + \bar{x}_2 + \ldots + \bar{x}_n\}$



If we have population data who is normally distributed. On this we applied Sampling distri. and after that calculating sample mean.

↳ when we ploted all those mean → We got normal distribution

② $X \not\simeq N(\mu, \sigma)$



$n \geq 30$

$S_1 \rightarrow \{x_1, x_2, \ldots x_{30}\} \rightarrow \bar{x}_1$

$S_2 \rightarrow \{x_2, x_3, \ldots, x_{30}\} \rightarrow \bar{x}_2$

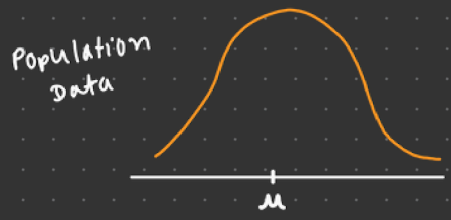
$\vdots$

$S_n \rightarrow \bar{x}_n$

⇒ CLT Properties

the central limit theorem says that the sampling dis. of the mean will be always be normally distributed, as long as the sample size is large enough. Regardless of wheather the population has a normal, poisson, binomial or any other distr. the sampling distribution of the mean will be normal.

Important for the interview



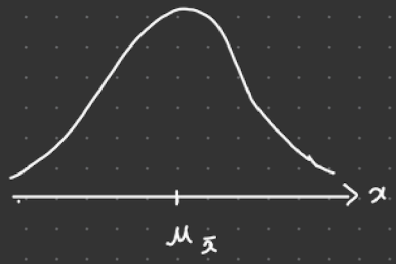
$x \simeq N(\mu, \sigma)$

$\sigma \Rightarrow$ population std

$\mu \Rightarrow$ Population mean

$n \Rightarrow$ Sample size.

Sampling distribution of mean (CLT)



$\bar{x} \simeq N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

n can be any value

Q What is the Std in sample distribution ?

↳ Std $\Rightarrow \frac{\sigma}{\sqrt{n}}$

Infrential Statistics

Estimate :- It is an observe numerical value used to estimate an unknown population parametter.

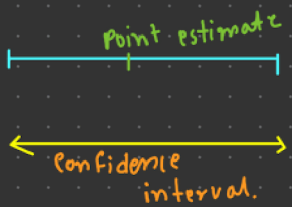
① **Point estimate** → Single numerical value used to estimate the unknown population perumetter.

* Sample mean is a point estimate of a population mean.

There is a huge gape, to counter those gape we use Interval estimate.

② **Interval estimate** → Range of value used to estimate the unknown population parametter.

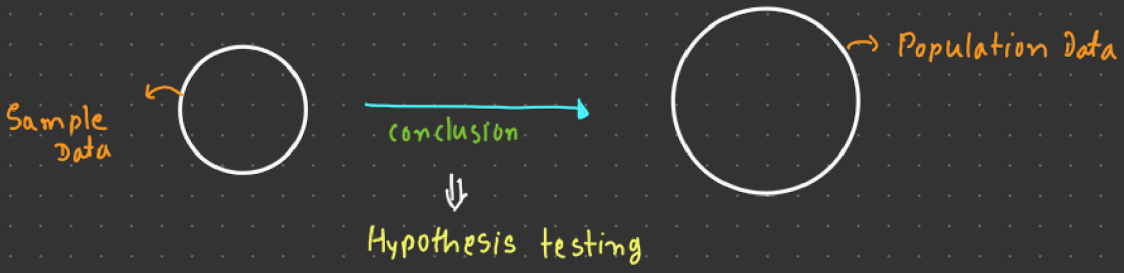
* Interval estimate of population parameter are called confidence interval.



55 - 65

A Hypothesis and Hypothesis testing mechanism :-

↳ In inferential Stats. used for → Conclusion or Infernces.



* Hypothesis testing Mechanism →

ex → Person 🦹 → done Crime

(1) Null Hypothesis (H_0) → The person is not Guilty

↳ The assumption that you are begining with.

(2) Alternate Hypothesis (H_1) → The person is Guilty

↳ Opposite of Null Hypothesis

(3) Experiments → profe collect (by help of → Statistical Analysis) (P-value)

↳ { DNA, Tests, finge prints }

(4) Accept the null hypothesis OR Reject the null hypothesis

For example, College at district a stats says its average passed percentage of statistic are 85%. A new college opened in the district and it was found that a sample of 100 Students have a pass % of 90 with a standard deviation of 40%. Does this school have a different Passed %.

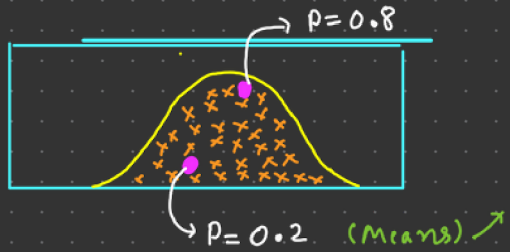
Ans → Null Hypothesis → (H_0) → $\mu = 85\%$

Alternate hypothesis → (H_0) → $\mu \neq 85\%$

P-Value

The p-value is a number, calculated from, a Statistical test, that discribe how likely you are to have found a particular set of observations if the null hypothe. were true. P-Values are used in hypothesis testing to help decide wheather to reject the null hypothesis.

Ex → Using keyboards space tap.



Out of 100 touches in this key the probability of touching in this regin ≥0

↳ Hypothesis testing

Ex → Wheather coin is fair or Not. {100 Tosses}

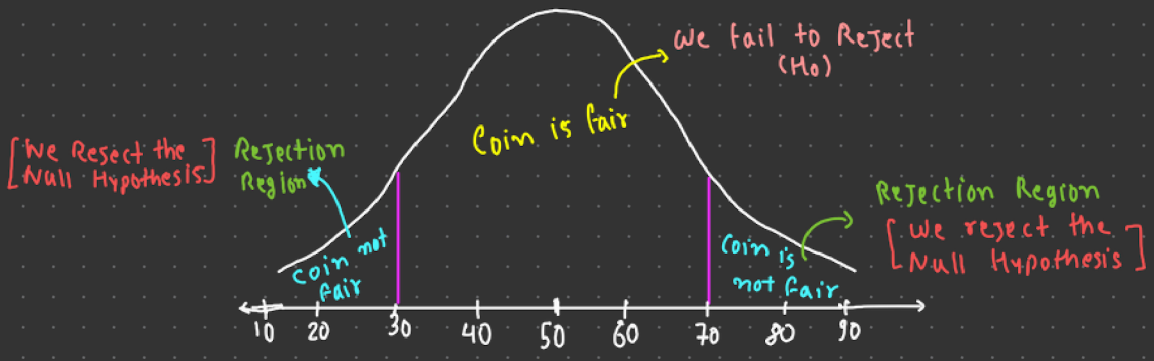
Ans →

① Null Hypothesis (H_0): Coin is fair

② Alternate hypothesis (H_1): Coin is not fair

③ Experiment :

$$\begin{cases} P(H) \to 0.5,\ P(T) \to 0.5 \\ P(H) \to 0.6,\ P(T) \to 0.4 \\ P(H) \to 0.7,\ P(T) \to 0.3 \end{cases}$$



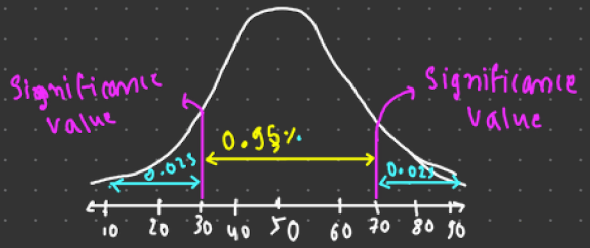
* Significance value (α)

Let, $\alpha = 0.05$

Confidence interval (C.I.)

C.I. $\Rightarrow 1 - 0.05$

$\Rightarrow 0.95$



* Conclusion

Let, $P = 0.01\%$

$\rightarrow$ $P <$ Significance

$\therefore$ We reject the null hypothesis

else:

we fail to reject null hypothesis.

Hypothesis testing and statistical Analysis :-

(A) Z - Test
(B) T - Test } Average
(C) CHI SQUARE } categorical data
(D) ANNOVA } varianie

(A) Z - test :-

Condition : Z-score only apply on where
(i) Population std
(ii) $n \geq 30$

Problem The average height of all the residents in a city is 168cm. with a $\sigma = 3.9$, A Doctor belive that mean to be different. He measure the height of 36 individuals and found the average height to be 169.5 cm

(a) State Null and alternate hypothesis.
(b) At 95% confidence level, is there enough evidence to reject the null hypothesis.

Method 1 ⟹ Z - test

$\mu = 168cm$, $\sigma = 3.9$, $n = 36$, $\bar{x} = 169.5cm$

(a) Null hypothesis (H_0) = $\mu = 168cm$

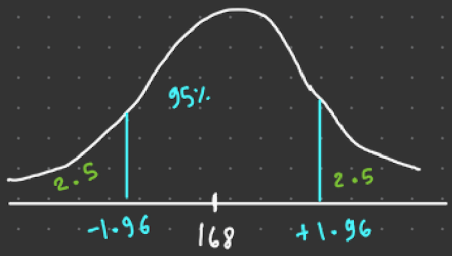
Alternate hypothesis (H_1) = $\mu \neq 168$ {2 Tail test}

It can be greater OR lesser then also.

(b) C.I. = 0.95 , $\alpha = 1 - 0.95$

$= 0.05$

Decision boundry



* Area under the curve

$1 - 0.95 \Rightarrow 0.05$

0.25 0.25

$\Rightarrow 1 - 0.25 \Rightarrow \boxed{0.9750}$

Area under the curve

Z-Score $\rightarrow 0.9750 \rightarrow +1.96$

Z-Score for population data

$$\text{Z-Score} = \frac{x_i - \mu}{\sigma}$$

Z-Score for sample data

$$\text{Z-Score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

* Statistical Analysis

$$\text{Z-Score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Rightarrow \frac{169.5 - 168}{3.9/\sqrt{36}} \Rightarrow 2.31$$

If Z-test value is less then -19.6 OR Greater then +1.96 we

Reject the Null hypothesis

else

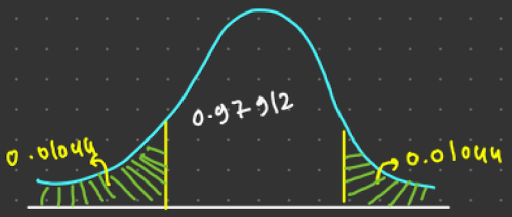
we Accept the Null hypothesis

$2.31 > +1.96$ { So, we Reject the null hypothesis }

Method 2 → p-value

* Statistical Analysis →

$$2\text{-Score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{169.5 - 168}{3.9/\sqrt{36}} \Rightarrow 2.31$$



In Z-table 2.31 ⇒ 0.98956

1 - Aera under the curve

↳ 1 - 0.98956

⇒ 0.01044

* How to calculate p-value →

⇒ 0.01044 + 0.01044

⇒ 0.02088

if P value < Significance

0.02088 < 0.05 {We Rejecting the null hypothesis}

else-

we accept the null hypothesis.

↳ Here we accepting the null hypothesis.

Problem: 2 A factory manufacture bulbs with a average warenty of 5 yr. with standard deviation of 0.50 yr. A worker belive that the bulb will manufacture in less then 5 yr. He test a sample of 40 bulbs and find the average time to be 4.8 years.

(a) Stats the null hypothesis and alternate

(b) At a 2% Significance level, is there enough evidence to support the idea that the warrenty should be revised.

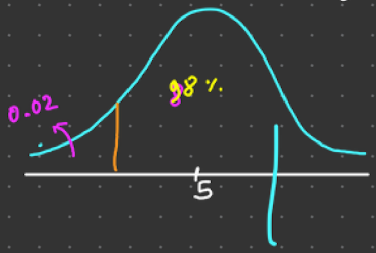
Ans → $\mu = 5$, $\sigma = 0.50$, $\bar{x} = 4.8$, $n = 40$

↳ (A) Null hypothesis (H_0) ⇒ $\mu = 5$

↳ Alternate hypothesis (H_1) ⇒ $\mu < 5$ {2 Tail test}
(Only checking one side)

* Method (p-Value)

- Decision Boundry

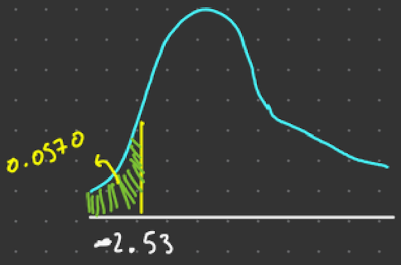


C.I. → 0.98

α = 1 − 0.98

⇒ 0.02

→ p-value → z-test → $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ → $\frac{4.8 - 5}{\frac{0.05}{\sqrt{40}}}$ → $\frac{-0.2}{0.079}$ → −2.53



↳ Area under the curve of −2.53, z value is = 0.0570

p-value = 0.0570

if

Pvalue < Significance

0.0570 < 0.02 → False

* Conclusion → The warranty needs to be revised. So, we accept the Null hypothesis.